Simple Linear Regression

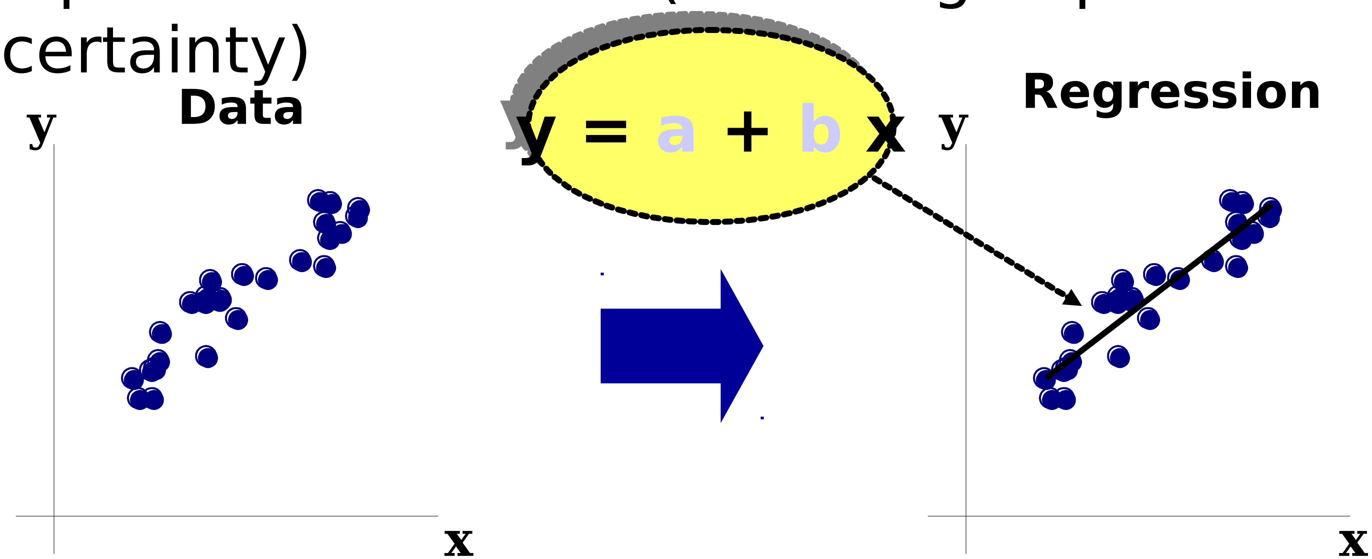
Chapter 7

Regression Analysis

- **A relationship between variables may exist due to 1 of 4 possible reasons:**
  - **Chance**
    - » **useless since this relationship can not be quantified**
  - **A relationship to a 3rd set of circumstances**
    - » **a more direct relationship is desired since it provides a better explanation of cost**
  - **A functional relationship**
    - » **a precise relationship that seldom exists in cost estimating**
  - **<u>A causal type of relationship</u>**

Definition of Regression

- Regression Analysis is used to describe a *statistical* relationship between variables
- Specifically, it is the process of estimating the "best fit" parameters of a specified function that relates a dependent variable to one or more independent variables (including implicit uncertainty)

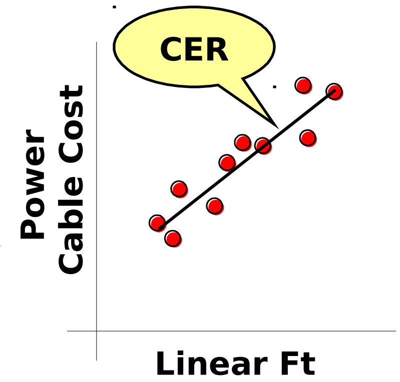


Regression Analysis in Cost Estimating

- **If the dependent variable is a cost, the regression equation is often referred to as a *Cost Estimating Relationship* or *CER***
  - **The independent variable in a CER is often called a *cost driver***

Examples of cost drivers:

Cost	Cost Driver (single)
Aircraft Design	# of Drawings
Software	Lines of Code
Power Cable	Linear Feet

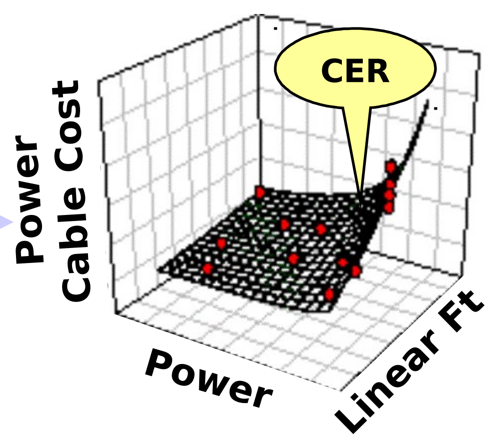


3

  - **A CER may have multiple cost drivers:**

Example with multiple cost drivers:

Cost	Cost Driver (multiple)
Power Cable	Linear Feet Power



Linear Regression Model

- Cost is the dependent (or unknown) variable; generally denoted by the symbol Y.
- The system's physical or performance characteristics form the model's known, or independent, variables which are generally denoted by the symbol X.
- The linear regression model takes the following form:

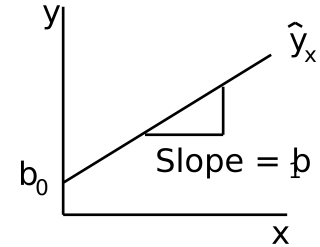
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

  where b_0 (the Y intercept) and b_1 (the slope of the regression line) are the unknown regression parameters and ε_i is a random error term.

- It is assumed that ε_I ~ N(0, σ2) and iid.

Linear Regression Model

- **We desire a model of the form:**



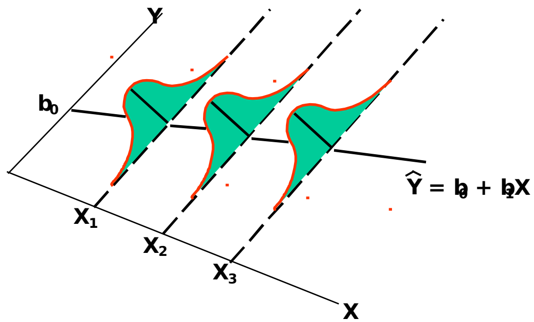
$$\hat{y}_x = b_0 + b_1 x$$

- **This model is estimated on the basis of historical data as:**

$$y_i = b_0 + b_1 x_i + e_i$$

where

$$e_i \sim N(0, \sigma_x^{\,2}), \text{ and iid}$$



- **b_1 and b_0 are chosen such that the sum of the squared residuals is minimized (Least Squares Best Fit).**

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \hat{y} = \text{residua}$$

$$\sum (y_i - \hat{y})^2 = \text{minimum}$$

Least Squares Best Fit (LSBF)

- **To find the values of b_0 and b_1 that minimizes** $\sum (y_i - \hat{y})^2$ **one may refer to the "Normal Equations."**

$$\sum Y = nb_0 + b_1 \sum X$$

$$\sum XY = b_0 \sum X + b_1 \sum X^2$$

- **With two equations and two unknowns, we can solve for b_0 and b_1.**

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \overline{Y} - b_1 \overline{X}$$

An Example

- **Suppose we're analyzing the production cost of radio comm sets.**
- **The average production cost of all radio comm sets in your data set is $250K**

$$\bar{Y} = \$250K$$

- ***Then* you develop an estimating relationship between production cost and radio comm set weight using LSBF.**

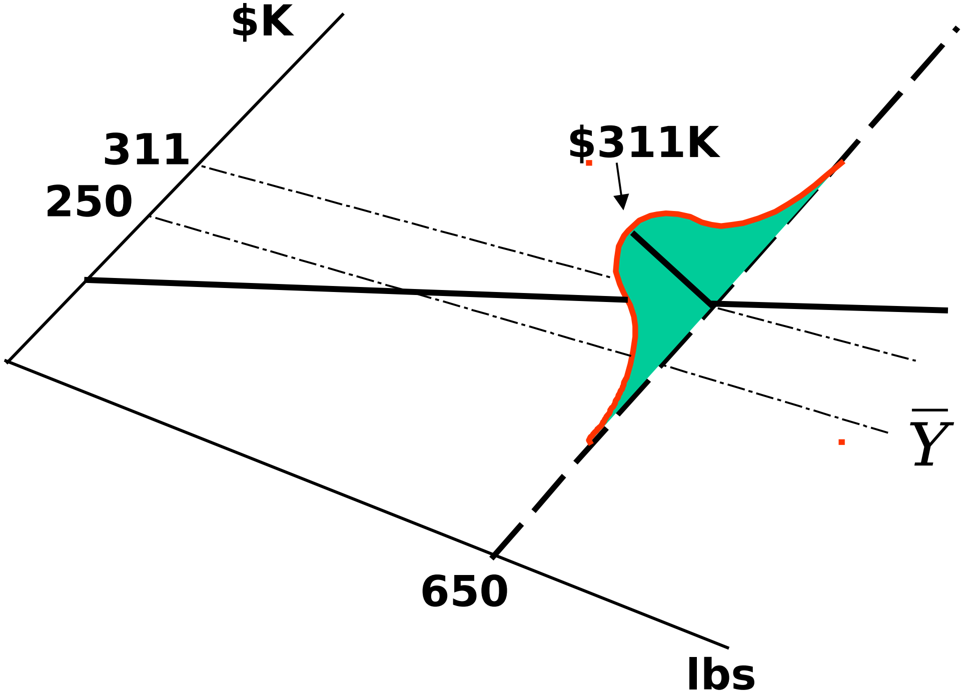
$$\hat{Y} = \$25K + \$0.44K(\text{weight in lbs})$$

- **Now you want to estimate the production cost of a 650 lb. radio comm set.**

$$\hat{Y} = \$25K + \$0.44K(650\text{ lbs}) = \$311K$$

An Example

- What do these numbers mean?
- $250K is the estimate of the average production cost of all radio comm sets in the population.
- $311K is the estimate of all radio comm sets in the population <u>that have a weight of 650 lbs.</u>



$$\hat{Y} = \$25K + \$0.44K(\text{weight}$$

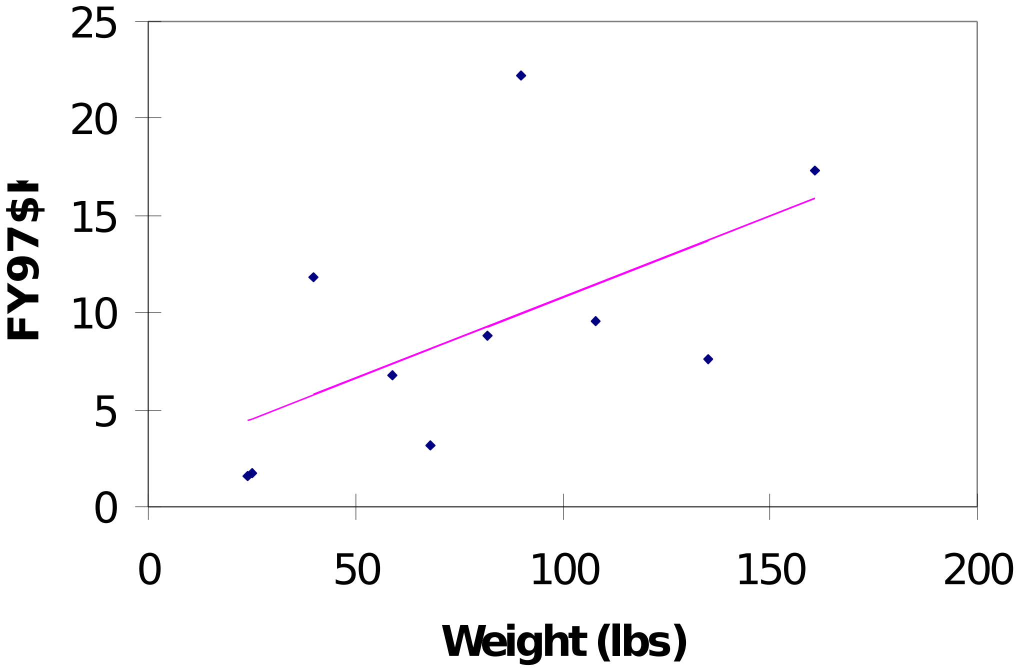
$$\overline{Y} = \$250K$$

Another Example

- **Recall the transmogrifier? Now lets look at the relationship between transmogrifier weight (lbs) and average unit production cost.**

Historic Transmogrifier
Average Unit Production Cost

System	FY97$K	Weight (lbs)
1	22.2	90
2	17.3	161
3	11.8	40
4	9.6	108
5	8.8	82
6	7.6	135
7	6.8	59
8	3.2	68
9	1.7	25
10	1.6	24



The Regression Model

- **The first time, we'll crank it out by hand...**

$$Y = \text{Average Unit Production Cost (FY97\$K)}$$

$$X = \text{Weight (lbs)}$$

$$n = 10$$

$$\overline{X} = \frac{\sum X_i}{n} = 79.2 \text{ lbs} \qquad \overline{Y} = \frac{\sum Y_i}{n} = \$9.06K$$

$$\sum XY = 8{,}739.4 \qquad \sum X^2 = 81{,}540$$

$$\hat{b}_1 = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2} = \frac{8{,}739.4 - (10)(79.2)(9.06)}{81{,}540 - (10)(79.2)^2} = 0.0831$$

$$\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X} = 9.06 - 0.0831(79.2) = 2.48$$

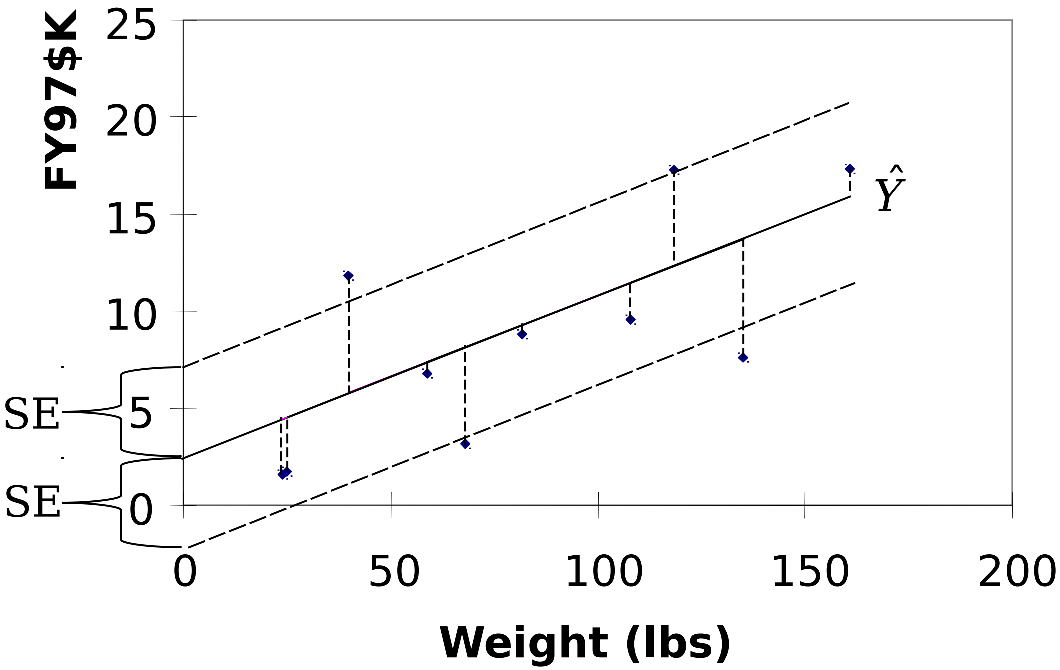
$$\therefore \hat{Y}_X = \$2.48K + (\$0.0831K)X$$

Standard Error

- **Standard Error =** $S_{\hat{y}}$ **= the standard deviation about the regression line. The smaller the better.**

$$SE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

n-k-1, where k is numb
of independent variabl

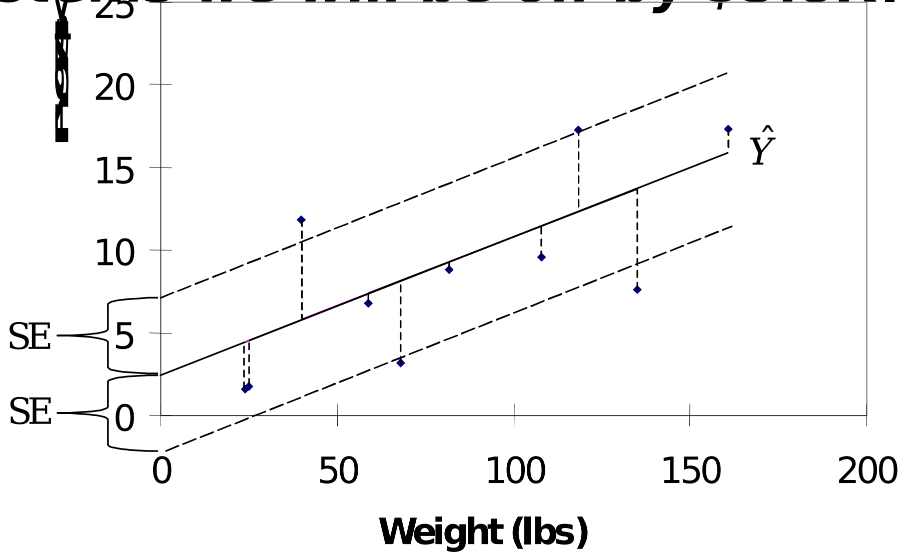


Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
22.2	10.0	12.2	149.87
17.3	15.9	1.4	2.07
11.8	5.8	6.0	35.98
9.6	11.5	-1.9	3.44
8.8	9.3	-0.5	0.24
7.6	13.7	-6.1	37.19
6.8	7.4	-0.6	0.34
3.2	8.1	-4.9	24.30
1.7	4.6	-2.9	8.15
1.6	4.5	-2.9	8.25
		Sum	**269.83**

Standard Error

$$SE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{269.83}{8}} = \$5.8K$$

- **For the transmogrifier data, the standard error is $5.8K.**
- **This means that *on "average" when predicting the cost of future systems we will be off by $5.8K*.**



Coefficient of Variation

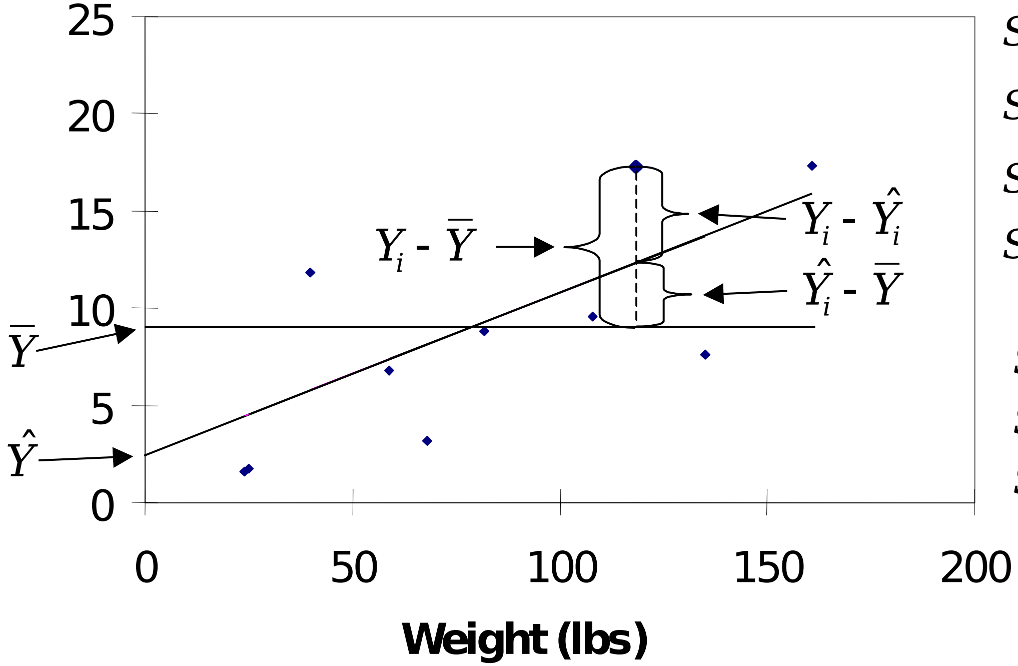
- **Coefficient of Variation (CV)**

$$CV = \frac{SE}{\overline{Y}} = \frac{\$5.8K}{\$9.06K} = 64\%$$

- **This says that *on "average", we'll be off by 64% when predicting the cost of future systems*. The smaller the better.**

Analysis of Variance

- **Analysis of Variance (ANOVA)**



$SST = \text{Total Sum of Squares} = \sum (Y_i - \bar{Y})^2$

$SSE = \text{Sum of Squared Errors} = \sum (Y_i - \hat{Y}_i)^2$

$SSR = \text{Sum of Squares Regression} = \sum (\hat{Y}_i - \bar{Y})^2$

$SST = SSE + SSR$

$SST = \text{Total Variation}$

$SSE = \text{Unexplained Variation}$

$SSR = \text{Variation explained by Regression}$

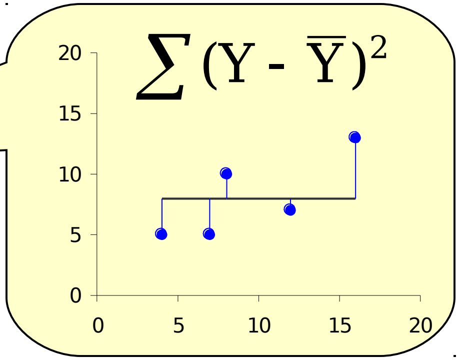
	df	*SS*	*MS = SS/df*	*F = MSR/MSE*	*Significance F P(b1=b2=0)*
SSR	1	130.00	130 (MSR)	3.85	0.0852
SSE	8	269.83	33.7 (MSE)		
SST	9	399.82			

Analysis of Variance (ANOVA)

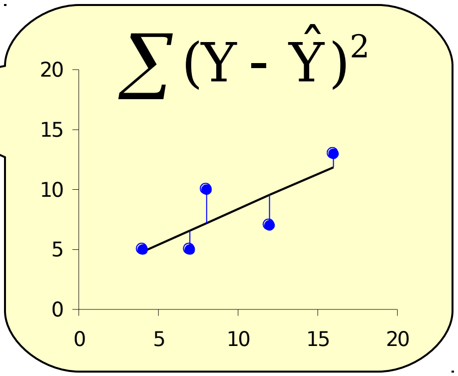
Measures of Variation

7

1. **Total Sum of Squares (SST):**
   The sum of the squared deviations **between the data and the average**

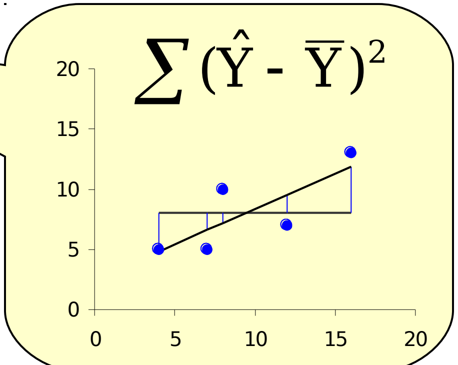


2. **Residual or Error Sum of Squares (SSE):**
   The sum of the squared deviations **between the data and the regression line**
   "The unexplained variation"



3. **Regression Sum of Squares (SSR):**
   The sum of the squared deviations **between th**
   the averag
   "The ex

SST = SSE + SSR
"total" = "unexplained" + "explained"



Analysis of Variance (ANOVA)

Mean Measures of Variation

– **Mean Squared Error (or Residual) (MSE):**

$$\mathbf{MSE} = \frac{\mathbf{SSE}}{\mathbf{n - k}}$$

– **Mean of Squares of the Regression (MSR):**

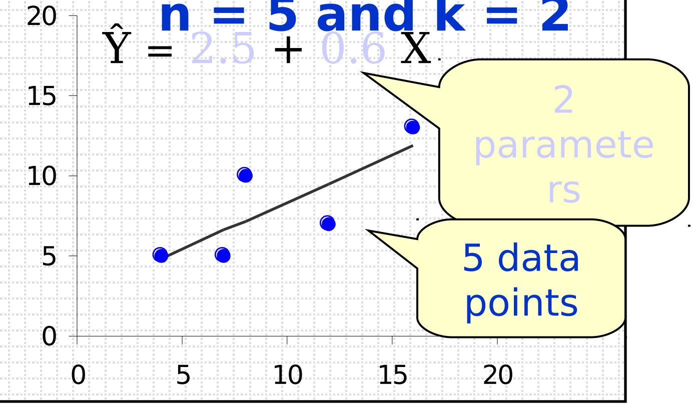
$$\mathbf{MSR} = \frac{\mathbf{SSR}}{\mathbf{k - 1}}$$

where:

n = # data points

k = # equation parameters

e.g. in our toy problem
n = 5 and k = 2



10

The denominator for each of the above is called the *degrees of freedom*, or *df*, associated with each type of variation

Coefficient of Determination

- **Coefficient of Determination (R^2) represents *the percentage of total variation explained by the regression model*. The larger the better.**

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2 = \frac{130}{399.8} = 1 - \frac{269.8}{399.8} = 0.3252 = 32.5\%$$

- **R^2 adjusted for degrees of freedom (Adj. R^2) takes into account the increased uncertainty due to a small sample size.**

$$R^2_{adj} = 1 - \frac{SSE / (n - (k+1))}{SST / (n-1)} = 1 - \frac{269.8 / 8}{399.8 / 9} = 0.2408 = 24.1\%$$

The t statistic

- **For a regression coefficient, the determination of statistical significance is based on a *t test***
  - **The test depends on the ratio of the coefficient's estimated value to its standard deviation, called a *t statistic***
- **This statistic tests the marginal contribution of the independent variable on the reduction of the unexplained variation.**
- **In other words, it tests the strength of the relationship between Y and X (or between Cost and Weight) by testing the strength of the coefficient b_1.**
- **Another way of looking at this is that the t-statistic tells us how many standard deviations the coefficient is from zero.**
- **The t-statistic is used to test the hypothesis that X and Y (or Cost and Weight) are NOT related at a given level of significance.**
- **If the test indicates that that X and Y <u>are</u> related, *then we say we prefer the model <u>with</u> b_1 to the model <u>without</u> b_1.***

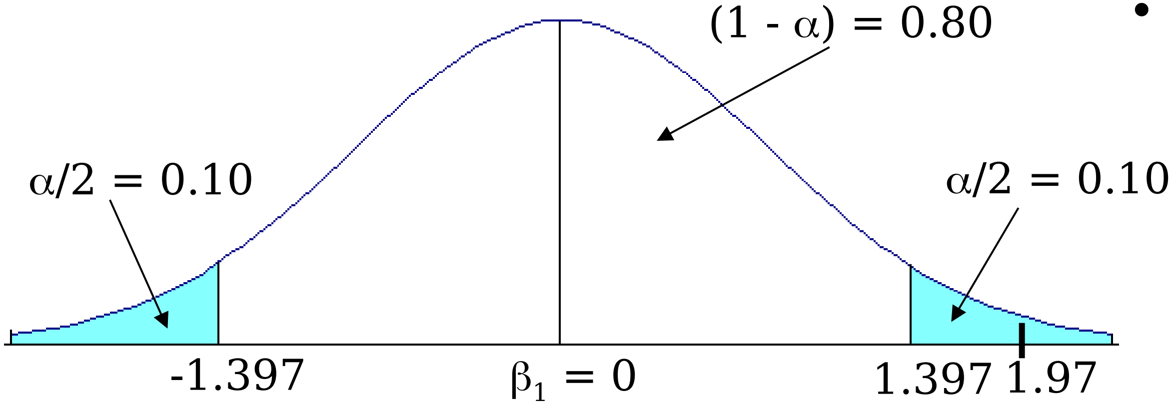
The t statistic

$H_0 : \beta_1 = 0$ (Cost is not related to weight) (prefer model without weight)

$H_a : \beta_1 \neq 0$ (Cost and weight are related) (prefer model with weight)

Test statistic: $t_{\beta_1} = \dfrac{b_1 - \overset{0}{\beta_1}}{s_{b_1}} = \dfrac{b_1}{s_{b_1}} = \dfrac{b_1}{SE \Big/ \sqrt{\sum (X - \overline{X})^2}} = \dfrac{0.0831}{5.8 \big/ 137.16} = 1.97$

- **Say we wish to test b_1 at the $\alpha = 0.20$ significance level. Refer to Table 6-2 with 8 degrees of freedom...**



- **Since our test statistic, 1.97, falls within the rejection region, we reject H_0 and conclude that we prefer the model <u>with</u> b_1 to the model <u>without</u> b_1.**

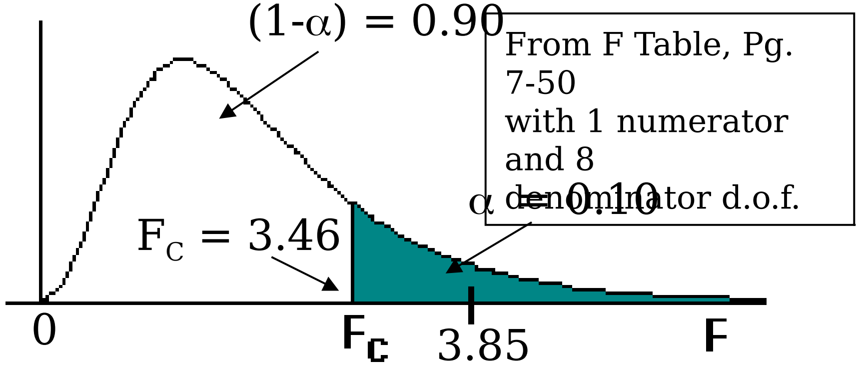
The F Statistic

- **The F statistic tells us whether the full model $\hat{Y}$ is preferred to the mean, $\bar{Y}$. That is, whether the coefficients of all the independent variables are zero**
- **Say we want to test the strength of the relationship between our model and Y at the α = 0.1 significance level...**

$$H_0: \beta_1 = \cdots = \beta_k = 0 \text{ (The model is invalid) (prefer } \bar{Y}\text{)}$$

$$H_a: H_0 \text{ is false (The model is valid) (prefer } \hat{Y}\text{)}$$

$$\text{Test statistic: } F = \frac{MSR}{MSE} = \frac{SSR/df_R}{SSE/df_E} = \frac{130/1}{269.8/8} = 3.85$$



- ***Since 3.85 falls within the rejection region, we reject H_0 and say** the full model is better than the mean as a predictor of cost.*

There's an Easier Way…

- **Linear Regression Results (Microsoft Excel):**

Regression Statistics	
Multiple R	0.5702
R Square	0.3251
Adjusted R Square	0.2408
Standard Error	5.8076
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	130.00	130.00	3.85	0.0852
Residual	8	269.83	33.73		
Total	9	399.82			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.477	3.823	0.648	0.535	-6.340	11.293
Weight (lbs)	0.083	0.042	1.963	0.085	-0.015	0.181

- **Now the information we need is seen at a glance.**

Important Results

- From the Excel Regression output we can glean the following important results:
  - R^2 or Adj. R^2: The bigger the better.
  - CV: Divide Standard Error by $\bar{Y}$ (calculated separately). The smaller the better.
  - Significance of F: If less than α then we prefer the model $\hat{Y}$ to the mean $\bar{Y}$. Else, vice versa.
  - P-value of coefficient b_1: If less than α then we prefer the model with b_1, else we prefer it without b_1.
- These statistics will be used to compare other linear models when more than one cost driver may exist.

Treatment of Outliers

- **In general, an outlier is a residual that falls greater than 2σ from $\bar{Y}$ or $\hat{Y}$.**
- **The standard residual is**

$$\frac{Y_i - \hat{Y}}{SE} \quad or \quad \frac{X_i - \bar{X}}{s_X} \quad or \quad \frac{Y_i - \bar{Y}}{s_Y}$$

- **Recall that since 95% of the population falls within 2σ of the mean, then in any given data set, we would expect 5% of the observations to be outliers.**
- **In general, do not throw them out unless they do not belong in your population.**

Outliers with respect to X

- All data should come from the same population. You should analyze your observations to ensure this is so.
- Observations that are so different that they do not qualify as a legitimate member of your independent variable population are called outliers with respect to the independent variable, X.
- To identify outliers with respect to X, simply calculate $\bar{X}$ and S_X. Those observations that fall greater than two standard deviations from $\bar{X}$ are likely candidates.
- You expect 5% of your observations to be outlier, therefore the fact that some of your observations are outliers is not necessarily a problem. You are simply identifying observations that warrant a closer investigation.

Example Analysis of Outliers with Respect to X

Rang	$\overline{X}$	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$\frac{(X_i - \overline{X})}{S_X}$
600	823	-223	49785	-0.59
925	823	102	10379	0.27
450	823	-373	139222	-0.99
420	823	-403	162510	-1.07
1000	823	177	31285	0.47
800	823	-23	535	-0.06
790	823	-33	1097	-0.09
1600	823	777	603535	2.06

S_X 377.65

Outliers with Respect to Y

- **There are two types of outliers with respect to the dependent variable.**
  - **Those with respect to Y itself.**
  - **Those with respect to the regression model, $\hat{Y}$.**
- **Outliers with respect to Y itself are treated in the same way as those with respect to X.**
- **Outliers with respect to $\hat{Y}$ are of particular concern, because those represent observations our model does not predict well.**
- **Outliers with respect to $\hat{Y}$ are identified by comparing the residuals to the standard error of the estimate (SE). This is referred to as the "standardized residual."**

$$\frac{(Y_i - \hat{Y})}{SE} = \#\,\mathrm{of\ Standard\ Error}$$

- **Outliers are those with residuals greater than ±2 std errors.**

Remedial Measures

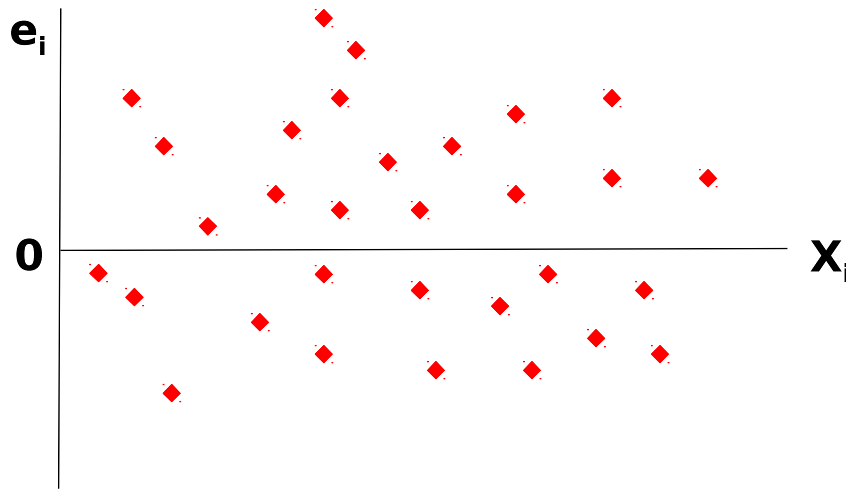
- Remember: the fact that you have outliers in your data set is not necessarily indicative of a problem. The trick is to determine WHY an observation is an outlier.
- Possible reasons why an observation is an outlier.
  - Random Error: No problem
  - Not a member of the same population: If so, you want to delete this observation from your data set.
  - You’ve omitted one or more other cost drivers.
  - Your model is improperly specified.
  - The data point was improperly measured (it’s just plain wrong).
  - Unusual event (war, natural disaster).
  - A normalization problem.

Remedial Measures

- **Your first reaction should not be to throw out the data point.**
- **Assuming the observation belongs in the sample, some options are:**
  - **Dampen or lessen the impact of the observation through a transformation of the dependent and or independent variables.**
  - **Develop two or more regression equations (with and without the outlier)**
- **Outliers should be treated as useful information.**

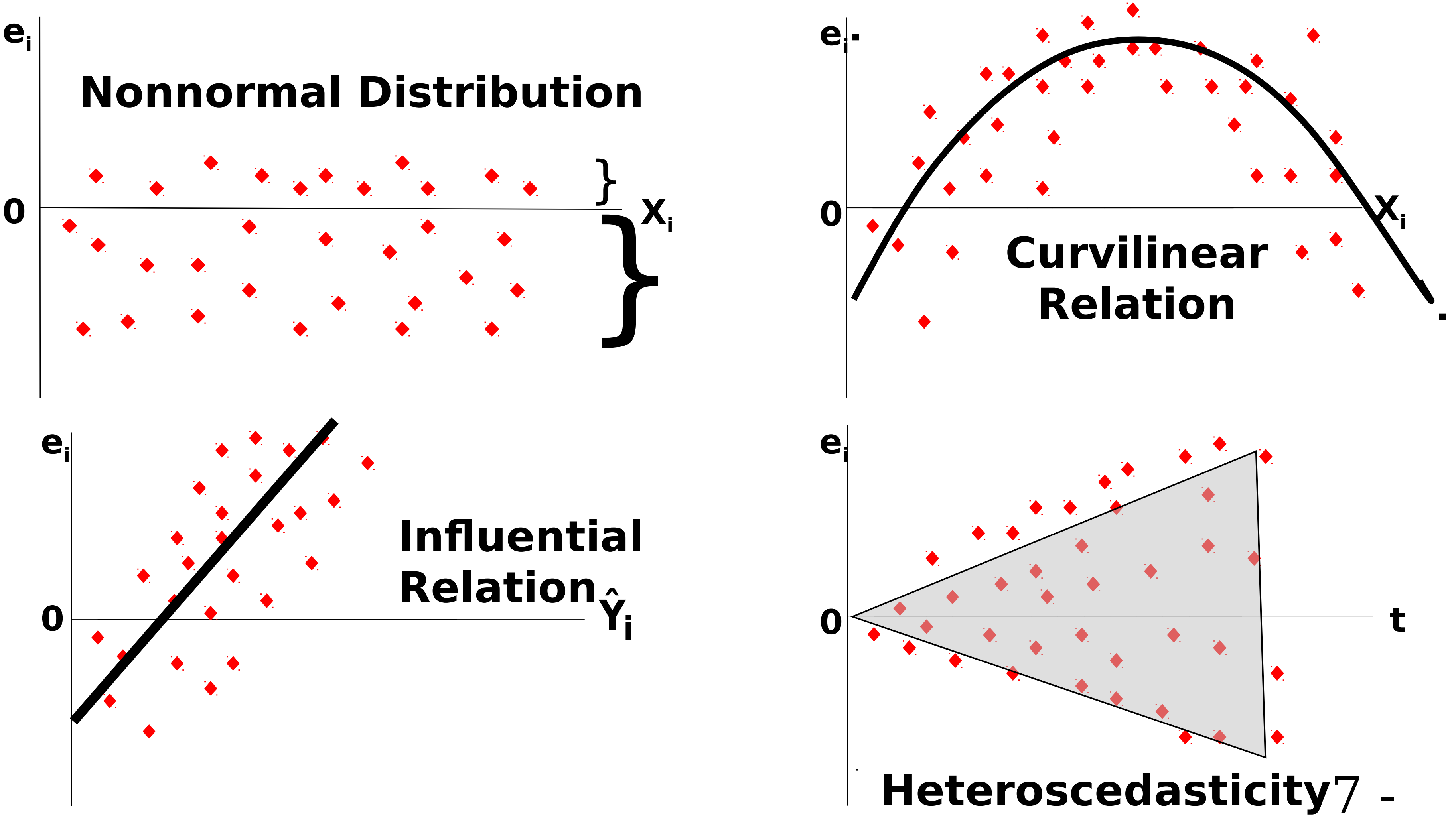
Model Diagnostics

- **If the fitted model is appropriate for the data, there will be no pattern apparent in the plot of the residuals versus X_i, $\hat{Y}_i$, etc.**
  - **Residuals spread uniformly across the range of X-axis values**



Model Diagnostics

- If the fitted model is not appropriate, a relationship between the X-axis values and the e_i values will be apparent.

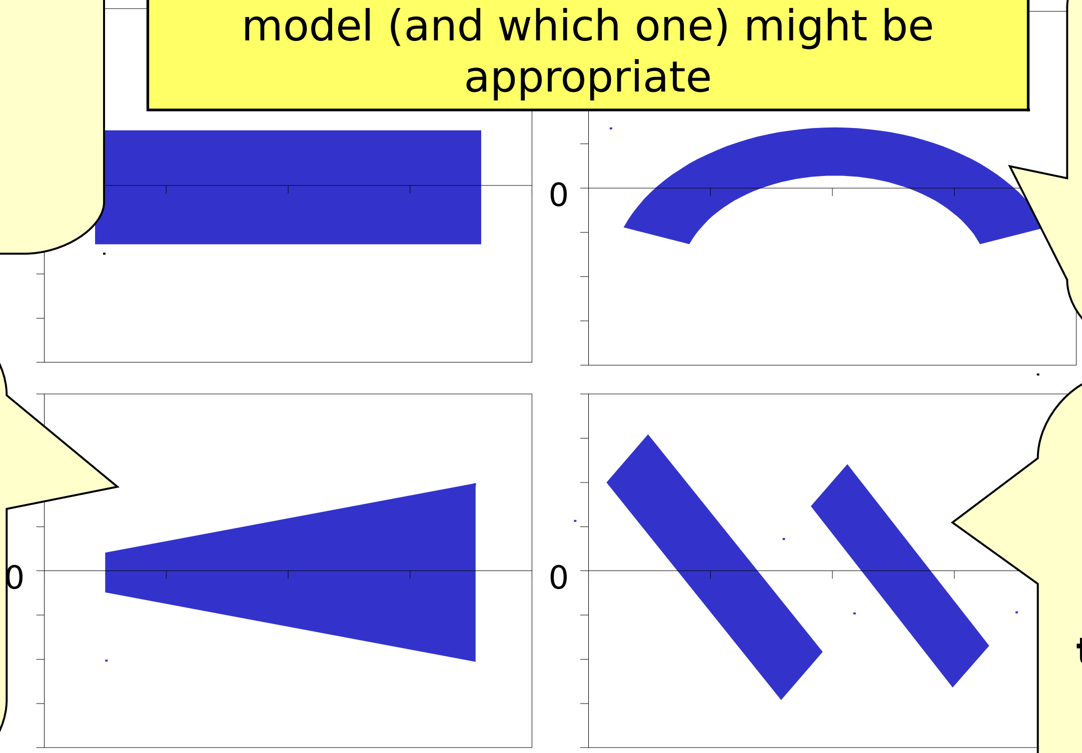


Example Residual Patterns

Tip: A residual plot is the primary way of indicating whether a non-linear model (and which one) might be appropriate

Good residual pattern:
- **Independent with x**
- **Constant variation**

Residuals not independent with x:
A curvilinear model is probably more appropriate in this case



Residuals do not have constant variation: Weighted Least Squares approach should be examined

Residuals not independent with x:
e.g., in learning curve analysis, this pattern might indicate loss of learning or injection of new work

Usually the residual plot provides enough vi insight to determine whether or not linear OL regression is appropriate. If the picture is inconclusive, statistical tests exist to help determine if the OLS assumptions hold[1].

Non-Linear Models

- **Data transformations should be tried when residual analysis indicates a non-linear trend**

  X´ = 1/X X′ = 1/Y X′ = log X Y′ = ln Y Y′ = log Y

  - **CER is often non-linear when independent variable is a performance parameter**

    $$Y = aX^{b}$$

    $$\log Y = \log a + b \log X \;\Rightarrow\; Y' = a' + bX'$$

    » **log-linear transform allows use of linear regression**

    » **predicted values for Y are “log dollars” which must be converted**

- **r^2 is potentially misleading when using a log model**

Other Concerns

- **When the regression results are illogical (i.e., cost varies inversely with a physical or performance parameter), omission of one or more important variables may have occurred or the variables being used may be interrelated**
  - **Does not necessarily invalidate a linear model**
  - **Additional analysis of the model is necessary to determine if additional independent variables should be incorporated or if consolidation/elimination of existing variables is required**

Assumptions of OLS

(1) Fixed X

- **Can obtain many random samples, each with the same X values but different Y_i values due to different e_i values**

(2) Errors have mean of 0

- **$E[e_i] = 0$**

(3) Errors have constant variance (homoscedasticity)

- **$Var[e_i] = \sigma^2$ for all I**

(4) Errors are uncorrelated

- **$Cov[e_i, e_j] = 0$ for all $i \neq j$**

(5) Errors are normally distributed

- **$e_i \sim N(0, \sigma^2)$**